

# **Statistical methods for detecting signals of natural selection in the wild**

MARKKU KARHUNEN

Department of Biosciences  
Faculty of Biological and Environmental Sciences  
University of Helsinki

ACADEMIC DISSERTATION

To be presented for public examination with the permission of the Faculty of Biological and Environmental Sciences of the University of Helsinki in Auditorium XV, (Unioninkatu 34) on October 18th at 12 o'clock noon.

HELSINKI 2013

SUPERVISED BY: Prof. Otso Ovaskainen  
University of Helsinki, Finland

REVIEWED BY: Prof. Mikko Sillanpää  
University of Oulu, Finland  
  
Dr. Jarrod Hadfield  
University of Oxford, Finland

EXAMINED BY: Prof. Jukka Corander  
University of Helsinki, Finland

CUSTOS: Prof. Veijo Kaitala  
University of Helsinki, Finland

ISBN 978-952-10-9195-7 (paperback)  
ISBN 978-952-10-9196-4 (PDF)  
<http://ethesis.helsinki.fi>

Yliopistopaino  
Helsinki 2013



# Contents

ABSTRACT	6
TIIVISTELMÄ	7
SUMMARY	9
<b>1. INTRODUCTION</b>	<b>9</b>
Evolutionary forces	9
Genetic architecture	10
Measuring phenotypic differentiation	12
Problems and challenges in $Q_{ST}$ - $F_{ST}$ comparisons	13
A word on statistical methods	14
<b>2. AIMS OF THE THESIS</b>	<b>15</b>
<b>3. MATERIAL AND METHODS</b>	<b>16</b>
Statistical techniques	16
Empirical data	17
<b>4. RESULTS AND DISCUSSION</b>	<b>17</b>
Development of methods	17
Software and real data	18
Conceptual limitations	19
<b>5. ACKNOWLEDGEMENTS</b>	<b>20</b>
<b>REFERENCES</b>	<b>21</b>
<b>I</b> Estimating Population-Level Coancestry Coefficients by an Admixture F Model	25
<b>II</b> A new method to uncover signatures of divergent and stabilizing selection in quantitative traits	45
<b>III</b> driftsel: an R package for detecting signals of natural selection in quantitative traits	77
<b>IV</b> Bringing habitat information into statistical tests of local adaption in quantitative traits: a case study of nine-spined sticklebacks	103

The thesis is based on the following articles, which are referred to in the text by their Roman numerals:

I

Karhunen M. and Ovaskainen O. 2012. Estimating population-level coancestry coefficients by an admixture F-model. – *Genetics* 192: 609-617.

II

Ovaskainen O., Karhunen M., Zheng C., Cano Arias J. M. and Merilä J. 2011. A new method to uncover signatures of divergent and stabilizing selection in quantitative traits. – *Genetics* 189: 621-632.

III

Karhunen M., Merilä J., Leinonen T., Cano Arias J. M. and Ovaskainen O. 2013. driftsel: an R package for detecting signals of natural selection in quantitative traits. – *Molecular Ecology Resources* 13: 746-754.

IV

Karhunen M., Ovaskainen O., Herczeg G. and Merilä J. 2013. Bringing habitat information into statistical tests of local adaption in quantitative traits: a case study of nine-spined sticklebacks. – accepted for *Evolution* subject to minor revision.

Table of contributions

	I	II	III	IV
Original idea	MK, OO	JM, JC, OO	OO	JM
Data collection / generation	MK *	MK	TL, JC, MK	GH
Mathematical derivations	MK, OO	OO, CZ	MK	MK, OO
Programming	MK, OO	OO, MK	MK	MK
Manuscript preparation	MK, OO	OO, JC, JM	MK, OO, JM	MK, JM

MK = Markku Karhunen, OO = Otso Ovaskainen, JM = Juha Merilä, JC = José Manuel Cano Arias, CZ = Chaozhi Zheng, TL = Tuomas Leinonen, GH = Gábor Herczeg  
\*In I, empirical data obtained from other research teams are used as detailed in Methods.

## ABSTRACT

Much of evolutionary biology attempts to explain why the phenotypes of local populations have diverged from a common ancestral type. It is often tempting to explain the observed pattern of differentiation by natural selection, but this often lacks scientific justification, because random genetic drift – i.e. the gradual change of allelic frequencies due to random sampling of alleles from parents to offspring – can cause striking phenotypic differentiation, being arguably a simpler and a more eloquent explanation.

Random genetic drift needs to be ruled out as an explanation to argue for a scenario of phenotypic differentiation based on natural selection. The rate of random drift in a study system can be measured by using neutral DNA markers, such as microsatellites. Subsequently, these measurements can be compared with the observed pattern of phenotypic differentiation. This is usually done by using so-called  $Q_{ST} - F_{ST}$  comparisons. However, these comparisons suffer from a number of conceptual limitations, and their power to detect signals of natural selection is relatively low. This thesis presents a re-thinking of the neutral DNA-phenotype comparison. To this

end, the Chapters of this thesis discuss the evolution of randomly drifting quantitative traits, resulting in a probability distribution over the evolutionary process. I use this distribution to construct statistical tests that can be used to detect signals of natural selection in quantitative traits.

The method of this thesis is derived from the first principles of quantitative genetics (assuming diploidism and additive genetic architecture), and it is developed into two R software packages (driftsel and RAFM), intended for users without extensive skills in Bayesian statistics. This thesis also investigates the performance of the methods and the interpretation of the model estimates obtained as a by-product of the neutrality tests. As it turns out, the method developed here can detect signals of natural selection even in cases where usual  $Q_{ST} - F_{ST}$  comparisons fail. In addition to the methodological contributions, the work included into this thesis has also an empirical dimension demonstrating occurrence of adaptive differentiation among three-spined (*Gasterosteus aculeatus*) and environmental adaptation in nine-spined stickleback (*Pungitius pungitius*) populations.

# TIIVISTELMÄ

Evoluutiobiologian yksi keskeisistä tavoitteista on selittää miksi ja miten paikallispopulaatioiden fenotyypit ovat eriytyneet yhteisestä alkumuodosta. On houkuttelevaa selittää havaittu eriytyminen aina luonnonvalinnan avulla, mutta tämä ei ole ilman lisäselvityksiä perusteltua, koska geneettinen satunnaisajautuminen, eli geenifrekvenssien vähittäinen muuttuminen satunnaisvaihtelun vuoksi, pystyy tuottamaan samantapaista eriytymistä. Satunnaisajautuminen on myös yksinkertaisempi ja siksi elegantimpi selitysmalli.

Geneettisen satunnaisajautumisen mahdollisuus on suljettava pois, jotta voidaan perustellusti väittää, että fenotyyppien eriytymisen johtuu luonnonvalinnasta. Satunnaisajautumisen nopeutta voidaan tarkastella neutraalien DNA-markkereiden, kuten mikrosatelliittien, avulla. Näin saatavia arvioita voidaan verrata havaittuun fenotyyppien vaihteluun. Tämä tehdään useimmiten niin sanottujen  $Q_{ST} - F_{ST}$ -testien avulla. Nämä testit kärsivät kuitenkin useista käsitteellisistä ongelmista, ja niiden tilastollinen voima on suhteellisen huono. Tämä väitöskirja esittää uuden ajatusmallinneutraalin DNA:n jafenotyyppien vertailua varten. Väitöskirjassa tarkastellaan kvantitatiivisten ominaisuuksien geneettistä satunnaisajautumista

ja johdetaan niiden todennäköisyysjakauma evolutiivisen prosessin suhteen. Tämän jakauman avulla rakennetaan tilastollisia testejä, joita voidaan käyttää luonnonvalinnan signaalien osoittamiseksi kvantitatiivisessa aineistossa.

Tämän väitöskirjan metodit johdettiin kvantitatiivisen genetiikan perusperiaatteista (diploidia ja additiivinen periytyminen), ja ne kehitettiin aina kahdeksi R-ohjelmistopakettiksi asti (driftsel ja RAFM). Nämä ohjelmistopaketit soveltuvat myös sellaisille käyttäjille, joilla ei ole laajoja tietoja bayesiläisistä tilastomenetelmistä. Lisäksi tässä työssä tarkasteltiin kehitettyjen menetelmien toimivuutta ja neutraliteettitestien sivutuotteena saatavien parametristimaattien tulkintaa. Osoittautuu, että tässä työssä kehitetyn metodin avulla pystytään osoittamaan luonnonvalinnan signaaleja sellaisissakin tapauksissa, joissa  $Q_{ST} - F_{ST}$ -testit eivät toimi. Metodologian lisäksi tämä väitöskirjatyö käsittää myös empiirisen komponentin. Työssä pystytään osoittamaan fenotyyppien adaptiivinen eriytyminen kahdella ekologisella mallilajilla, kolmipiikeillä (*Gasterosteus aculeatus*) ja kymmenpiikeillä (*Pungitius pungitius*).





# SUMMARY

Markku Karhunen

*Metapopulation Research Group, Department of Biosciences, PO Box 65 (Viikinkaari 1), 00014 University of Helsinki, Finland*

## 1. INTRODUCTION

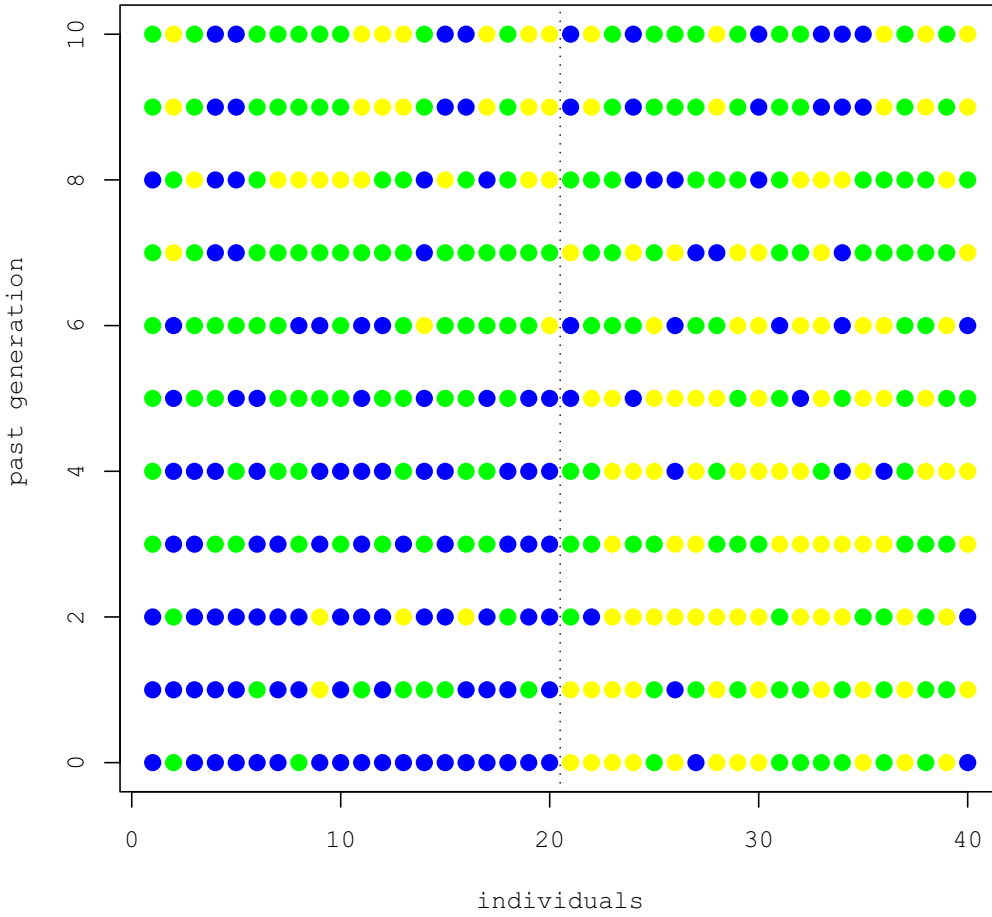
### EVOLUTIONARY FORCES

It is common wisdom that evolution occurs by the survival of the fittest: The fittest individuals, i.e. the ones best adapted to their environment, leave more descendants to future generations than the less fit ones. If the fitness differences are caused by individual differences in the phenotypes, and if this phenotypic variation has a heritable basis, the offspring generation tends to be better adapted to its local habitat than the parental generation (Darwin 1859). This process is known as evolution by natural selection. It has been demonstrated over and over again in different natural and experimental set-ups that this mechanism produces local adaptations. Classical examples include Grant's & Grant's work with Darwin's finches (1995) and Reznick et al.'s work with guppies (e.g. Torres-Dowdall *et al.* 2012). Roff (1997) reviews a large number of experiments with quantitative traits.

However, not all evolution occurs by natural selection. Consider for example Fig. 1 which demonstrates the evolution of a Mendelian phenotype in two small, isolated populations. Population 1 is gradually becoming bluer, while population 2 is becoming yellower. This is however not a result of natural selection, because all phenotypes have been specified as equally fit in the simulation behind Fig. 1. What is then the cause of differentiation between these two populations? Mathematically, it is a sampling procedure where  $2N$  gene copies for offspring are sampled among the  $2N$  gene copies of parents. This is sampling with replacement, and prone to produce changes in gene frequencies. Thus we must conclude that mere chance can produce patterns such as Fig. 1. This phenomenon is known as random genetic drift, and its effects have been investigated by the so-called neutral (Kimura 1968, 1983) and nearly neutral (Ohta 1973, 1992) theories of molecular evolution. In Fig. 1, we in fact see a clear example of neutral theory, as the phenotype is directly equivalent to the diploid genotype of the AB locus.

How does random genetic drift relate to more complex phenotypes such as behavioral, morphological or life-history traits? These are often polygenic: they are not usually affected by a single underlying locus, but by a number of different genes, each with a relatively small effect (e.g. Flint & MacKay 2009; Hill 2010; MacKay 2004). A number of questions arise here. Firstly, how do the different genes interact, i.e. how is the phenotype determined as a function of the multi-locus genotype? Secondly, how often do changes in the gene frequencies of causative loci occur due to random drift, and are these changes frequent enough to produce phenotypic differentiation in quantitative traits, as opposed to natural selection? Thirdly, can the effects of random drift and natural selection be distinguished in experimental data? These questions turn out to be more complex than they seem, and remain still largely open. Proximal answers exist (e.g. Martin et al. 2008; Merilä & Crnokrak 2001; Whitlock 1999 and this thesis), but they are based on simplifying assumptions.

Before discussing the roles of natural selection and random drift in more detail, it should be noted that these two are not the only mechanisms that can change the genotypic composition of natural populations. Traditionally, two other evolutionary forces have been distinguished, namely mutation and migration (Ridley 2004). Mutation as such cannot produce large-scale genetic changes, but it is the ultimate source of all genetic variation which natural selection uses to create new phenotypic adaptations. Likewise, combined with a high rate of random genetic drift, even a mildly deleterious mutation may become fixed (Ohta 1973). The likelihood of such changes is the subject of neutral (Kimura 1983) and nearly neutral (Ohta 1973) theories. Thus, mutation produces variation, but the fate of new mutations is dictated by the interplay of random genetic drift and natural selection.



**Figure 1. Random genetic drift.** A simulated example of Mendelian evolution in two small isolated populations. There are  $N = 10$  hermaphroditic individuals per generation per population, and mating occurs at random. Initially, the populations split from a panmictic ancestral population, and they evolve for  $T = 10$  generations. Blue = genotype AA, green = AB, yellow = BB. The changes in genetic composition are entirely due to random drift.

In principle, migration could produce large-scale genetic changes. As a theoretical example, a small local population could be overrun by an influx of maladapted migrants, which represents an evolutionary change into the direction of reduced fitness. However, such changes would not be maintained for long, unless the influx of migrants persists, as in some source-sink systems (Dias & Blondel 1996; Pulliam & Danielson 1991). In practice, migration is often seen as counterpart of random drift: Migration typically decreases the quantitative measures of random genetic drift (Rousset 2004; Whitlock 2011; and see below), so that populations that interchange migrants are said to be ‘drifting less’. Consequently, methods that aim to

distinguish signatures of natural selection and random genetic drift in data should also distinguish between natural selection and migration.

### GENETIC ARCHITECTURE

Let us now focus on the first question of this Introduction, i.e. the interplay between different genes coding for a quantitative trait. The simplest model of polygenic inheritance is the so-called additive model (see Lynch & Walsh 1998) where the additive genetic effect of an individual is defined as

$$a := \sum_{j=1}^n (g_{j1} + g_{j2}) \quad (\text{Eq. 1})$$

where the sum ranges over  $n$  genomic loci, and  $g_{j1}$  and  $g_{j2}$  denote haploid genotypes of the first and second gene copies of locus  $j$ , respectively. (Here we assume that the alleles are labeled according to their additive effects; hence we can use  $g_{j1}$  and  $g_{j2}$  in place of the allele identifiers in Eqs. 2-4 below.) The additive effect  $a$  is often known as the breeding value. Diploidism is implicitly assumed in Eq. 1. Also, it should be noted that the breeding value is not necessarily the same as the observable phenotype, as the rearing environment also affects many traits (Eq. 5 and below).

In addition to the additive component described above, the total genotypic value can be further decomposed into dominance and epistatic components (Lynch & Walsh 1998). Dominance and epistasis imply non-linear interactions among haploid gene copies. They can be defined as follows. In absence of epistatic effects, the total genotypic value is assumed to be

$$y = \sum_{j=1}^n f_j(g_{j1}, g_{j2}) \quad (\text{Eq. 2})$$

where  $f_j(g_{j1}, g_{j2})$  indicates the total effect of the diploid genotype of locus  $j$ . Using this, the dominance effect can simply be defined as

$$d := y - a = \sum_{j=1}^n [f_j(g_{j1}, g_{j2}) - (g_{j1} + g_{j2})]. \quad (\text{Eq. 3})$$

The total effect of genotypes is still linear across loci in Eq. 3, because the non-linearity occurs within loci. This simplification can be removed by a generalization which includes epistasis, which can be defined generally as a departure from the additive-dominance model:

$$\varepsilon := z - (a + d) = z - \sum_{j=1}^n f_j(g_{j1}, g_{j2}), \quad (\text{Eq. 4})$$

$z$  implying the total effect of genotype. The epistatic effects  $\varepsilon$  can be further decomposed into e.g. additive-additive and additive-dominance components by calculating similar sums and differences over loci, but these considerations are sufficient for our purposes. Regarding the environmental influences, a common option is to define the phenotype as a sum of genotype and an environmental effect,

$$p = z + e. \quad (\text{Eq. 5})$$

Admittedly, this too is a simplification, as it does not allow for the genotype-environment interaction (colloquially known as  $G \times E$  interaction or  $G \times E$  correlation). Alas, in full generality, the concept of genetic architecture can only be described by the vague notion that the phenotype is a stochastic function of genotype and environment. The focus of this thesis is not on that level of generality, but on workable mathematical and statistical models. Hence, I adopt Eq. 1 and Eq. 5 as the model of genetic architecture, and study the conclusions that can be drawn within this framework using mathematical methods. This is also the starting point of many previous related methods (e.g. Whitlock 1999, 2008), and according to some authors at least, a realistic model of genetic architecture (Hill *et al.* 2008).

## MEASURING RANDOM GENETIC DRIFT

Regarding the second main theme of this Introduction, namely influence of random genetic drift on causative loci, let us first discuss the concept of random genetic drift in more detail. A natural starting point is to ask how the gene frequency of a single locus varies due to random sampling in repeated matings. Denoting generation by  $t = 1, 2, \dots$ , for one locus with gene frequency  $\pi_t$  and a randomly mating diploid population of size  $N$ , the variance of gene frequency in offspring is

$$\text{Var}(\pi_{t+1} | \pi_t) = \frac{\pi_t(1 - \pi_t)}{2N} \quad (\text{Eq. 6})$$

which follows from elementary probability calculus. The variance of gene frequency over  $T$  generations is

$$\text{Var}(\pi_{t+T} | \pi_t) = \pi_t(1 - \pi_t) \left[ 1 - \left( 1 - \frac{1}{2N} \right)^T \right] \quad (\text{Eq. 7})$$

which can be shown by using quite elementary techniques (see e.g. Templeton 2006). However, both Eqs. 6 and 7 depend on the initial gene frequency  $\pi_t$ , which is arguably a property of the locus, not of the breeding system. Thus arises the concept of Wright's (1951)  $F_{ST}$ :

$$F_{ST} = \frac{\text{Var}(\pi_{t+T} | \pi_t)}{\pi_t(1 - \pi_t)}, \quad (\text{Eq. 8})$$

which is the evolutionary variance of gene frequency scaled by the initial variation represented by  $\pi_i(1-\pi_i)$ . To date,  $F_{ST}$  is still one of the most common statistics in population genetics (Whitlock 2011), perhaps due to its easy interpretation in many classical population models (Rousset 2002), such as the island model and the stepping-stone model (see Hartl & Clark 2007), or a single isolated population (Eqs. 6-8).  $F_{ST}$  can also be generalized for a heterogeneous group of local populations and used as a summary statistic of genetic differentiation among them (Rousset 2002, 2004; Slatkin 1991). The effect of migration is incorporated in  $F_{ST}$ : In a range of different demographic models, it can be shown that migration decreases the value of  $F_{ST}$ , and isolation increases  $F_{ST}$  (Rousset 2004). This is also my conclusion in Chapter I. As stated above, migration can be seen as a counterpart of random genetic drift.

However, use of  $F_{ST}$  as a measure of random drift is hindered by the fact that all of quantities in Eq. 8 are typically unknown: The ancestral allele frequency  $\pi_i$  is not observed, because the ancestors are typically unknown, and hence also  $\text{Var}(\pi_{i+T} | \pi_i)$  is unknown. Moreover, estimating  $\text{Var}(\pi_{i+T} | \pi_i)$  is an impossible task, if only a single population and a single locus are considered, as this would correspond to estimating the variance from a single observation. Thus, it becomes necessary to combine information across different loci (e.g. Weir & Hill 2002), notwithstanding their different initial levels of polymorphism. Finally, the realized allele frequencies  $\pi_{i+T}$  on present generation are typically also unknown and need to be estimated from a finite subsample of individuals, thus generating additional sampling error.

### MEASURING PHENOTYPIC DIFFERENTIATION

Given the central role of  $F_{ST}$  in population genetics (Rousset 2002), it is natural to ask what kind of quantitative phenotypic differentiation is to be expected for a given level of  $F_{ST}$ . This can be answered by deriving a similar summary statistic for quantitative traits. Spitze (1993) defined  $Q_{ST}$  as

$$Q_{ST} = \frac{\sigma_B^2}{\sigma_B^2 + 2\sigma_W^2} \quad (\text{Eq. 9})$$

where  $\sigma_B^2$  is the variance of breeding values (Eq. 1) between populations, and  $\sigma_W^2$  is the variance of

breeding values within populations. Note that  $Q_{ST}$  is defined on basis of the (in principle) unobservable breeding values, not the observable phenotypes (Eq. 5): The analog of  $Q_{ST}$  for phenotypes is the  $P_{ST}$  (e.g. Brommer 2011). From Eq. 9 one immediately notes that  $Q_{ST}$  increases in  $\sigma_B^2$  and decreases in  $\sigma_W^2$ , which coincides with the intuitive concept of quantitative-genetic divergence – but why this index, and not some other, e.g.  $\sigma_B^2 / \sigma_W^2$ ? One reason is that  $Q_{ST}$  of Eq. 9 is expected to be equal to  $F_{ST}$  for a selectively neutral trait (Whitlock 1999). The argument of Whitlock (1999) is based on the following type of reasoning.

Let us assume an additive genetic framework (Eq. 1) where mutation increases the variance of breeding values by  $\sigma_m^2$  per generation. For deme  $i$  (whether a local population, species or metapopulation), the expected amount of additive genetic variance is

$$\sigma_i^2 = t_i \sigma_m^2 \quad (\text{Eq. 10})$$

where  $t_i$  is the average number of generations since the most recent common ancestor for two individuals in deme  $i$ , also known as the coalescence time. (This follows from a simple summation of independent random events.) From Eqs. 9 and 10, and a formula given by Wright (1969), Whitlock (1999) concludes that

$$Q_{ST} \approx \frac{t_B - t_W}{t_B} \quad (\text{Eq. 11})$$

where  $t_B$  is the average coalescence time in the whole metapopulation and  $t_W$  is the coalescence time within local populations. On the other hand, as Whitlock (1999) notes, Slatkin (1991) has shown (by simply taking the limit) that  $F_{ST}$  tends to Eq. 11 when mutation rate tends to zero.

Now we have two indices which are expected to be equal to each other in large populations, on the limit of a low mutation rate:  $F_{ST}$  that measures random genetic drift, and  $Q_{ST}$  which measures divergence of a quantitative trait. Comparing their values can produce useful information on the type of evolution that is operating on a quantitative trait (Merilä & Crnokrak 2001; Whitlock 2008; Whitlock & Guillaume 2009). Leinonen et al. (2008) survey 62 studies that do such comparisons. Their findings indicate that in 70 % of cases, the  $Q_{ST}$  values exceed their associated  $F_{ST}$  values. This could result from two things: Either diversifying natural selection is relatively common in nature, or the  $Q_{ST} - F_{ST}$  comparisons have a

**Table 1. Methods to detect natural selection (adapted from Leinonen *et al.* 2013).**

	Strengths	Weaknesses
$Q_{ST} - F_{ST}$ comparison (e.g. Merilä& Crnokrak 2001)	<ul style="list-style-type: none"> <li>- a widely used and understood method</li> <li>- signature of selection shows rapidly in <math>Q_{ST}</math> (tens rather than hundreds of generations)</li> </ul>	<ul style="list-style-type: none"> <li>- logistically difficult, requires common-garden experiments for many populations</li> <li>- a number of conceptual weaknesses (see text)</li> </ul>
$P_{ST} - F_{ST}$ comparison: calculating index of phenotypic differentiation from the wild (see Brommer 2011)	logistically simple, no common-garden experiment	observed differentiation in $P_{ST}$ may be due to phenotypic plasticity
multivariate comparisons: testing hypotheses such as $\mathbf{D} \approx \frac{2F_{ST}}{1 - F_{ST}} \mathbf{G}$ (e.g. Martin <i>et al.</i> 2008)	combining information from many traits and their genetic correlations results in increased power	essentially the same weaknesses as in $Q_{ST} - F_{ST}$ comparisons
genome scans for selected loci: e.g. comparing $F_{ST}$ values between loci and detecting outliers (i.a. Egea <i>et al.</i> 2008; Foll& Gaggiotti 2008; Storz 2005)	<ul style="list-style-type: none"> <li>- logistically simple, no common-garden experiment</li> <li>- tells something about the genetic basis of phenotypic variation</li> </ul>	<ul style="list-style-type: none"> <li>- allele frequencies at individual loci react slowly to natural selection (Le Corré&amp; Kremer 2012)</li> <li>- pronounced adaptive differentiation in phenotypes can pass unnoticed</li> </ul>
direct measurement of selection gradients: regressing survival and/or offspring number to observed characters (e.g. Lande& Arnold 1983)	direct observation of natural selection	<ul style="list-style-type: none"> <li>- assumptions on quantities that are not usually available (e.g. mutation rate, time since divergence)</li> <li>- current and past selection pressures could be different</li> </ul>

systematic bias, which could result from some of the technical challenges which I will discuss in the sequel. Finally, it should be noted that  $Q_{ST} - F_{ST}$  comparisons are by no means the best or only method to detect signals of natural selection among the background noise of random genetic drift (see Table 1). However, they are often the method of choice for evolutionary biologists, and they serve to illustrate an important principle: comparing divergence of phenotypes (e.g. measured by  $Q_{ST}$ ) and an independent measure of random drift (e.g.  $F_{ST}$ ). A similar comparison is also the fundamental idea of this thesis.

#### PROBLEMS AND CHALLENGES IN $Q_{ST}$ - $F_{ST}$ COMPARISONS

The derivation of Whitlock (1999) shows that  $Q_{ST}$  and  $F_{ST}$  are expected to have the same value for a

neutral trait and neutral DNA markers under certain conditions; if they are not equal, a departure from these conditions is indicated. On the other hand, if they are equal, it does not imply that these conditions are met. Mathematically, the assumptions of Whitlock (1999) are sufficient, but not necessary conditions for  $Q_{ST} \approx F_{ST}$  to hold. Biologically,  $Q_{ST} \approx F_{ST}$  does not imply absence of natural selection, only that its effect cannot be filtered apart from that of random genetic drift. On the other hand,  $Q_{ST} \neq F_{ST}$  can imply either presence of natural selection or some other departure from the assumptions used in Whitlock's (1999) derivation. Thus we see that the relationship of  $Q_{ST}$  and  $F_{ST}$  is only proximal, as the derivation is based on a number of simplifying assumptions.

Perhaps most importantly, and unlike usually assumed,  $Q_{ST}$  is a random variable. This is because Eq. 9

involves within and between-population components of phenotypic variance which are, in turn, also random variables subject to evolutionary stochasticity, i.e. random genetic drift (Miller *et al.* 2008; Whitlock 2008). In principle, it is the expectations of  $\sigma_B^2$  and  $\sigma_W^2$ , and not the expectation of  $Q_{ST}$ , that should be compared to  $F_{ST}$  for evolutionary inference. However, estimating these expectations from a small number of local populations is a statistical challenge, if not an impossible task (O'Hara & Merilä 2005; Whitlock & Guillaume 2009). As regarding the  $Q_{ST} - F_{ST}$  comparisons, ignoring the random nature of  $Q_{ST}$  could lead into the use of inflated point estimates (Miller *et al.* 2008), which in turn could result in the fact that the  $Q_{ST}$  values are, on average, higher than the  $F_{ST}$  values (Leinonen *et al.* 2008).

Furthermore, the relationship of  $Q_{ST}$  and  $F_{ST}$  has been derived on the limit of a low mutation rate (Whitlock 1999). While it would seem plausible, that this relationship is conserved if the mutation rate is the same for coding and neutral loci (influencing  $Q_{ST}$  and  $F_{ST}$ , respectively), this is not necessarily the case. The effect of mutation rate on  $Q_{ST}$  depends on the genetic architecture, namely the effect of mutation on the allelic values of the genetic variants (Kronholm *et al.* 2010). Secondly, it is known that microsatellite markers, routinely used for estimating  $F_{ST}$  in ecological applications, have a very high mutation rate, presumably much higher than the coding loci (Schlötterer 2000). Thus, concerns over the total effect of mutation rate on the  $Q_{ST} - F_{ST}$  comparisons has arisen (e.g. Edelaar *et al.* 2011).

The point that  $Q_{ST} \approx F_{ST}$  does not imply absence of natural selection raises the question, if the power of the test could be improved. What is needed is some measure of phenotypic divergence (in place of  $Q_{ST}$ ) to be compared with some measure of random genetic drift (in place of  $F_{ST}$ ). Ideally, these measures should account for a number of subtle phenomena, such as mutation and evolutionary stochasticity, even if assuming additive genetic architecture (Eqs. 1 and 5). Moreover,  $Q_{ST}$  as such (Eq. 9) has been derived for a scalar trait. A natural generalization of variance for multivariate data is the covariance matrix (Feller 1950). Hence, one may ask if an index such as Eq. 9 could be calculated for covariance matrices. This turns out to be the case (see Table 1). In fact, the additive genetic variance covariance matrices ( $\mathbf{G}$  matrices) have been

subject to vigorous research in evolutionary biology (Arnold *et al.* 2008; Chenoweth & Blows 2008; Lande 1979; Ovaskainen *et al.* 2008; Stepan *et al.* 2002), because they determine the most likely trajectories of evolutionary change (Lande 1979). On the other hand, they also change due to random drift and natural selection (Arnold *et al.* 2008; Stepan *et al.* 2002). Hence, comparing these matrices between populations could yield valuable information regarding the type of evolution in a study system (Chenoweth & Blows 2008; Martin *et al.* 2008; Ovaskainen *et al.* 2008).

Finally, also  $F_{ST}$  is a summary statistic, both in its traditional (Weir & Cockerham 1984; Wright 1951) and more modern (Rousset 2002, 2004; Slatkin 1991) definitions. Elaborate schemes are needed to average it over subpopulations, loci and individuals to yield reliable estimates (Weir & Hill 2002). This also implies that  $F_{ST}$  can have the same value for many types of metapopulation systems (Rousset 2004). For example, a system with one isolated population and two interconnected populations can have the same value of  $F_{ST}$  as a system with three moderately interconnected populations. Yet, one would expect to see different patterns of phenotypic differentiation due to random drift in these two systems, a fact which I will take into account in this thesis.

## A WORD ON STATISTICAL METHODS

At this point, a comment should be made on the statistical techniques used for this thesis. Statistics involves a measure of uncertainty or randomness. Mathematically, randomness is a well-defined concept yielding all the way from the Kolmogorov axioms (1933), but it is the biological interpretation of randomness that is obscure here. What does it mean when one says that  $Q_{ST}$  is a random variable? In this thesis, there are two kinds of randomness: process uncertainty and sampling variation in the data. Process uncertainty derives from the fact that one understands the evolutionary history of local populations as a random process. In fact, the effects of all four evolutionary forces are best seen as probabilistic: For example, natural selection implies that fit individuals have a high probability of breeding or survival, but the survival of the fittest is seldom certain. Sampling variation, on the other hand, stems from the fact that we cannot observe all individuals and molecular markers, so that data represent necessarily a random



sample among them. Thus, we see that  $Q_{ST}$  is a random variable regarding process uncertainty, and any estimate of  $Q_{ST}$  is also random with respect of sampling variation.

We need to quantify the effects of both sampling variation and process uncertainty to assess parameter uncertainty. In this thesis, I adopt the Bayesian paradigm (see e.g. Gelman *et al.* 2004) where the model parameters are random variables and ‘dual’ with the data, so that one may talk about their prior and posterior distributions. An illustrative example is perhaps in place. Let us assume that the ultimate parameter of interest is the ancestral frequency  $\pi_i$  of an arbitrary allele A, and yet we observe only a sample of genotypes on generation  $t+T$  which has the unobserved allele frequency  $\pi_{t+T}$ . It is quite straightforward to write the sampling model of the genotypes as a function of  $\pi_{t+T}$ . Typically the number of A alleles in the sample is distributed as

$$n_A \sim \text{Bin}(2n, \pi_{t+T}) \quad (\text{Eq.12})$$

where  $n$  is the sample size. The binomial distribution of Eq. 12 quantifies sampling variation, whereas the evolutionary process from  $\pi_i$  to  $\pi_{t+T}$  (so-called Wright-Fisher model, see Nicholson *et al.* 2002) represents process uncertainty, which can be quantified by some probability distribution  $p(\pi_{t+T}|\pi_i)$ . To obtain the posterior distribution for  $\pi_i$ , one needs to write the likelihood of data as a function of  $\pi_i$ . Symbolically,

$$p(n_A|\pi_i) = \int_0^1 \binom{2n}{n_A} \pi_{t+T}^{n_A} (1-\pi_{t+T})^{2n-n_A} p(\pi_{t+T}|\pi_i) d\pi_{t+T} \quad (\text{Eq. 13})$$

where  $p$  denotes probability density. This is known colloquially as integrating over  $\pi_{t+T}$ . In practice, this process may involve analytical techniques, such as integration, or numerical techniques, such as Markov chain Monte Carlo (MCMC, see Gelman *et al.* 2004). In complex models, the analytical techniques are infeasible, and the challenge lies in developing workable MCMC schemes or other numerical methods, such as Approximate Bayesian Computation algorithms.

In Chapter II, even more attention is paid on different types of randomness. Namely, four random processes

are distinguished: flow of alleles from the ancestral population to the presently observed local populations ( $F_g$ ), field sampling ( $F_s$ ), flow of alleles from the sampled individuals to the laboratory population ( $F_e$ ) and environmental effects of the laboratory individuals ( $F_c$ ). The distinction of these four processes turns out to be useful in the mathematical derivations of Chapter II, but for the purposes of this Summary, it is sufficient to discuss only process uncertainty and sampling variation as above.

## 2. AIMS OF THE THESIS

My objective is to develop a systematic framework for detecting natural selection in traits that have an additive genetic architecture (Eqs. 1 and 5). While many types and regimes of natural selection can be envisaged (for a quantitative treatment on a subset of models, see Kingsolver *et al.* 2001; Lande& Arnold 1983), I define neutrality as the starting point. Under neutrality, the genetic composition of populations changes due to mutation, migration or random drift, or some combination of these. Consequently, if we know what type of phenotypic change is attributable to each of these forces, we may interpret different deviations from this pattern as different signals of natural selection, two of which I will focus on in this thesis. The idea is the very same as in classical hypothesis testing: Consider for example the sample mean  $\bar{x}$  of independent normal observations  $x_i \sim N(\mu, \sigma^2); i=1, \dots, n$  (which is a textbook example). The expected pattern for  $\bar{x}$  is the sampling distribution

$$\bar{x} \sim N(\mu, \sigma^2 / n). \quad (\text{Eq.14})$$

Unusually high or low values of  $\bar{x}$  are interpreted as

deviations from the assumed pattern. For example,  $|\bar{x} - \mu| \geq 1.96\sqrt{\sigma^2 / n}$  is a deviation that occurs in only 5 % of cases under Eq. 14, and thus, it could be interpreted as a sign of departure from the null model. In analogy to this simplistic example, I attempt to develop test statistics that have a similar interpretation regarding the more complex model of evolution in absence of natural selection. What is needed here is a parameterization of the biological reality. Parameters are needed to quantify the effects of the remaining three evolutionary forces (*viz.* mutation, migration and drift), and presumably also the initial state of the study system prior to the action of these forces.

The above agenda concerns process uncertainty (see Introduction, *A word on statistical methods*). The parameters that are needed cover a wide range of demographic and genetic processes, which are typically unknown in the wild and partly uncontrollable also in experimental set-ups. Thus arises the need to estimate the values of these parameters from empirical data, and sampling variation comes into play. Ideally, a rigorous treatment of the model would allow integrating over both types of randomness, whether numerically or analytically, yielding posterior distributions for all parameters and test statistics. In this thesis, I aim to develop an automated algorithm which performs this, i.e. calculates the distributions and performs the statistical tests (Chapters I and II). Furthermore, I intend to develop this algorithm into a black-box-type computer program for users who do not have extensive knowledge of statistical computation (Chapters I, III and IV).

The output of the intended software can be defined to be the test statistics which measure deviations from the neutral pattern. The input must be defined as the empirical data, but what type of data? For  $Q_{ST} - F_{ST}$  comparisons (see Introduction), two kinds of data are needed: phenotypic data from quantitative traits, and molecular data from neutral marker loci. Concerning phenotypic data, it is less laborious to collect data without doing a common-garden experiment, but in this case, it is hard to separate genetic adaptation from phenotypic plasticity, i.e. to separate  $a$  and  $p$  (Eqs. 1 and 5). Consequently, common-garden experiments are often used in evolutionary biology when patterns of genetic differentiation are looked for (Merilä & Crnokrak 2001; Whitlock & Guillaume 2009). With respect of this, I define data collected for  $Q_{ST} - F_{ST}$  comparisons, i.e. common-garden data and molecular marker data, as the input for my computer program, with the possibility to include also habitat information, i.e. environmental covariates of interest (Chapter IV).

### 3. MATERIAL AND METHODS

#### STATISTICAL TECHNIQUES

This thesis aims at quantifying the null distribution of certain test statistics under neutrality, i.e. in absence of natural selection. To this end, analytical techniques

are employed as described in Results and Discussion, the Chapters of this thesis and their Appendices. To sample from the joint posterior distribution of the ultimate parameters, and consequently the test statistics, MCMC techniques are used. My work builds mainly on the Metropolis-Hastings algorithm (MH algorithm, see Gelman *et al.* 2004) where each parameter  $\theta$ , whether a scalar, vector or matrix, is updated according to following recipe.

1. Suppose the current value is  $\theta$ . Propose a new value  $\theta'$  from a probability distribution  $\pi(\theta \rightarrow \theta')$  which can depend on  $\theta$ .
2. Accept the new value by the probability
 
$$\xi(\theta \rightarrow \theta') = \min \left\{ 1, \frac{\pi(\theta'|\dots)\pi(\theta' \rightarrow \theta)}{\pi(\theta|\dots)\pi(\theta \rightarrow \theta')} \right\} \in [0,1]$$
 where  $\pi(\theta|\dots)$  denotes the full conditional (i.e. conditional on data and other parameters) probability density of  $\theta$ . Otherwise keep the current value.
3. Return to step 1.

This algorithm can be shown to converge to the posterior distribution of  $\theta$  (Hastings 1970). In practice, much of the work lies in inventing workable proposal distributions  $\xi(\theta \rightarrow \theta')$  for different types of  $\theta$  and implementing them with due numerical diligence. For example, in Chapter I, truncated Dirichlet distributions are needed to update allele frequencies in evolutionary lineages. In Chapters II-III, Wishart distributions are used to sample from the space of positive-definite matrices. (Covariance matrices such as  $\mathbf{G}$  matrices need to be positive definite, so that proposing other types of matrices would be inefficient, and accepting them would be fallacious.)

In this thesis, the variations of the MH algorithm are implemented in the R language (R Development Core Team 2012), and they are wrapped into R packages which are a standard format to distribute software among R users. This choice can be justified by two arguments: R is an open-source environment which is rapidly gaining popularity (Muenchen 2013), so that the method will probably be widely accessible. Secondly, R is a high-level programming language, which enables one to focus relatively more on the



conceptual, rather than the numerical side of statistical modeling.

## EMPIRICAL DATA

Empirical data are used here mainly as illustrative examples, apart from Chapter IV, where the data have a fundamental biological interest. In Chapter I, two data sets are used to demonstrate measurement of random genetic drift from microsatellite DNA (see Selkoe & Toonen 2006): data from common shrews (*Sorex araneus*) on the islands of Lake Sysmä (62°40'N, 31°20'E), earlier introduced by Hanski and Kuitunen (1986), and data from a set of four local populations of nine-spined sticklebacks (*Pungitius pungitius*) in Fennoscandia (which is a subset of data earlier introduced by Shikano *et al.* 2010).

In Chapter II, only simulated data are used. In Chapter III, a previously unpublished data set from three-spined sticklebacks (*Gasterosteus aculeatus*) is introduced and used to demonstrate the capabilities of the method, as implemented in R in this Chapter. In Chapter IV, the *Pungitius pungitius* data from II are combined with phenotypic measurements. In this study system, it has been so far difficult to show signals of selection by using rigorous statistical techniques, so that the inference on evolutionary history has been chiefly based on qualitative, yet logically justified, arguments (Herczeg *et al.* 2009; Merilä 2013). In Chapter IV, I employ the newly developed statistical techniques of Chapters I–III on this study system to accumulate new biological knowledge.

## 4. RESULTS AND DISCUSSION

### DEVELOPMENT OF METHODS

The work done in this thesis is illustrated by Figure 2. This thesis lies heavily on Chapter II which derives the distribution of population means for multivariate additive genotypes under random drift and migration. This is the desired parameterization of biological reality, and consequently, a description of process uncertainty. It turns out that the population means  $\mathbf{a}^p$  have the multivariate normal distribution

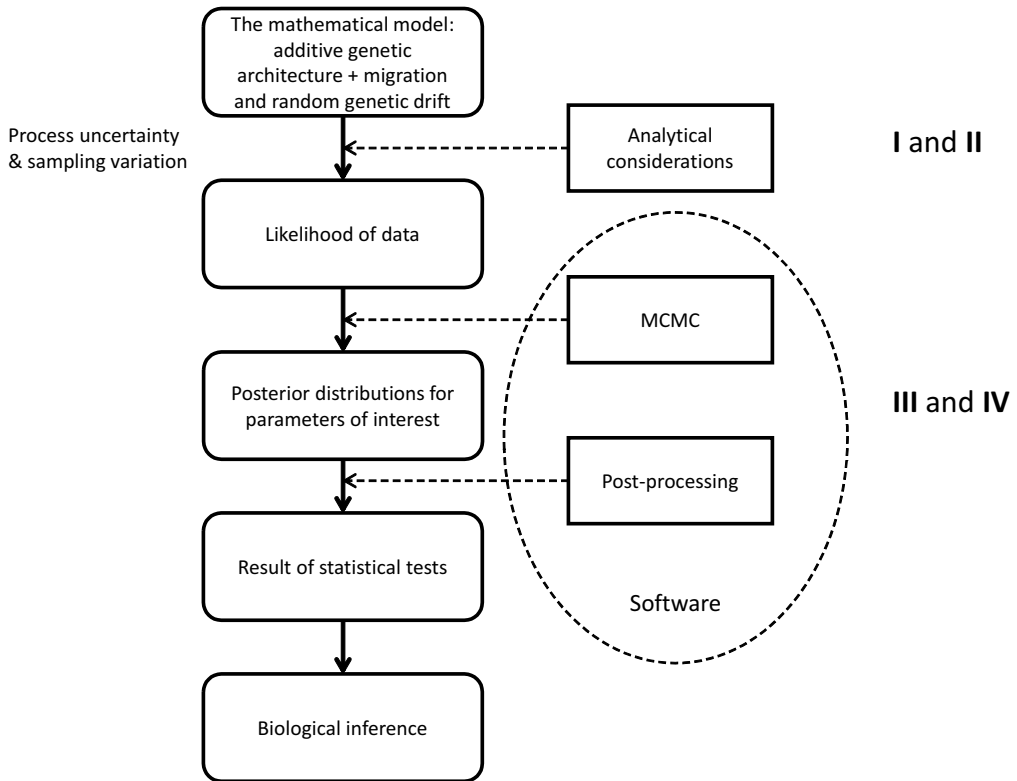
$$\mathbf{a}^p \sim \text{MVN}(\boldsymbol{\mu} \otimes \mathbf{I}, 2\mathbf{G}^A \otimes \boldsymbol{\theta}^p) \quad (\text{Eq.15})$$

where  $\boldsymbol{\mu}$  is the mean additive genotype of the ancestral population,  $\mathbf{I}$  is a unit vector,  $\mathbf{G}^A$  is the ancestral

$\mathbf{G}$  matrix,  $\otimes$  is a Kronecker product, and  $\boldsymbol{\theta}^p$  is the matrix of population-level coancestry coefficients. This is the analogy of Eq. 14 for multivariate random genetic drift. An analogy of Eq. 15 is well known at the level of individuals (standard animal model; see Lynch & Walsh 1998), but somewhat surprisingly, the population-level equation (i.e. Eq. 15) has, to my knowledge, not been presented before. Chapter II also introduces a scheme which enables estimating these parameters from data and introduces a test statistic  $S$  which measures the deviation of the observed population means from the distribution of Eq. 15.

Chapter I is a more detailed study on one of the components included in Chapter II, namely the admixture F-model intended for use with neutral DNA. This model permits estimating  $\boldsymbol{\theta}^p$  from molecular marker data. Matrix  $\boldsymbol{\theta}^p$  is the ultimate measure of random genetic drift and migration in this thesis (see Eq. 15). It would be desirable to write the likelihood of neutral DNA directly as a function of  $\boldsymbol{\theta}^p$ . However, we cannot integrate over process uncertainty here: We do not know, how the allele frequencies of multiallelic loci (such as microsatellites, Schlötterer 2000) change in finite populations due to random genetic drift. Solutions have been obtained for diffusion approximations of this process (Tavaré 1984), but these are computationally costly and thus difficult to implement in an MCMC sampling scheme. We model the allele frequencies as a mixture of Dirichlet distributions. These distributions are quantified by two demographic parameters,  $\boldsymbol{a}$  and  $\boldsymbol{\kappa}$ , which are common for all markers. As shown in Chapter I,  $\boldsymbol{\theta}^p$  is obtained as a function of these two, and  $F_{ST}$  as a function of  $\boldsymbol{\theta}^p$ . Theoretically, this model assumes that the local populations derive from evolutionary independent lineages which mix with each other one generation before the present time. This is a great and practically never met simplification, but preliminary simulations in Chapter I show that estimates of  $\boldsymbol{a}$  and  $\boldsymbol{\kappa}$  correspond to underlying demographic processes also in more complex situations with continuous gene flow,  $\boldsymbol{a}$  measuring random drift and  $\boldsymbol{\kappa}$  measuring migration.

Chapter III is chiefly a software paper demonstrating the estimation of the parameters introduced in Chapters I and II. In Chapter III, two extensions are made on the model of II: calculation of  $S$  is based on an exact formula, in place of a simulation procedure, and the likelihood of phenotypic data is extended



**Figure 2. The conceptual synopsis of this thesis.** This figure depicts one way to categorize the work done for this thesis. Chapters I and II are theoretically orientated, focusing on genotypic (I) and phenotypic (II) evolution in absence of natural selection. Chapters III and IV are more practically orientated, presenting software, statistical tests and visualizations. The distinction is not unambiguous, as Chapter I introduces the R package RAFM, and Chapter III and IV on the other hand also include minor revisions of the likelihood of Chapter II. Furthermore, the MCMC sampling scheme of driftsel was already used for the analysis of simulated data in Chapter II.

for a fully general study design (cf. full-sib design in Chapter II). Chapter IV is a mainly biological application of the method developed in Chapters I–III. In Chapter IV, the sampling model of Chapters I–III is extended for binary traits, which demands a new MCMC sampler. As a new innovation, test statistic  $H$  is now introduced. Test statistic  $S$  of II compares the (posterior distribution of) population means  $\mathbf{a}^P$  to their neutral distribution (Eq. 15). The new statistic  $H$  takes into account the environmental covariates, so that it compares the correlation of  $\mathbf{a}^P$  and covariates to that expected under Eq. 15. In other words,  $S$  asks if local populations are phenotypically more similar or dissimilar than expected, and  $H$  asks whether the phenotypes and environment correlate more, than would be expected on basis of their evolutionary history. Notably,  $H$  treats the environment as a space

of continuous variables, unlike methods that group local populations into discrete regions or habitat types (Chapuis *et al.* 2008; Whitlock & Gilbert 2012).

## SOFTWARE AND REAL DATA

The methods developed in this thesis are implemented as R the packages RAFM (presented in I) and driftsel (introduced in III). Both of these software packages are distributed on CRAN network in the standard tarball format familiar to R users (R Development Core Team 2012). Version 2.0 of driftsel covers the extensions made in Chapter IV, and it is presently available at <http://www.helsinki.fi/biosci/egru/software>. In addition to the estimation algorithms and calculation of the statistical tests, both versions of driftsel comprise two visualization functions: one that

explores the pattern of neutral genetic differentiation, and one that visualizes the posterior mean estimates of  $\mathbf{a}^p$ ,  $\boldsymbol{\mu}$ ,  $\mathbf{G}^A$  and  $\boldsymbol{\theta}^p$  in a space of two quantitative traits chosen by the user.

To conclude, these programs provide a biologically justified framework to detect signals of natural selection. This framework has been derived from first principles regarding migration and random drift in natural populations (I and II), and it has been operationalized as ready-to-use software. Case studies with simulated data demonstrate the power of this method and the qualitative conclusions that can be drawn from the results. Concerning empirical applications, the results are promising. In case of the nine-spined sticklebacks in Chapter IV, I am able to report signals of natural selection for both morphological and behavioural traits ( $S = 1.00, H = 1.00$  and  $S = 0.89, H = 0.99$ , respectively) – which has been anticipated for a long time (Merilä 2013), but not done in the rigorous sense.

### CONCEPTUAL LIMITATIONS

One obvious statistical limitation of the method developed in this thesis is that I focus on additive genetic architecture (Eq. 1). It is plausible that the distribution which forms the basis of the statistical tests (Eq. 15) would be conserved under random drift also in presence of non-additive effects. However, ignoring non-additive effects is known to lead into bias in the estimation of additive effects (Lynch & Walsh 1998), and thus the test statistics calculated by using these estimates are also biased. At present, driftsel does not include dominance (Eq. 3) or epistasis (Eq. 4). Furthermore, the neutral distribution (Eq. 15) has been derived assuming that both environmental ( $e$  in Eq. 5) and additive genetic effects ( $a$  in Eq. 1) have a multivariate normal distribution in the ancestral population. Obviously, this does not apply for many kinds of traits, such as integer-valued traits. In Chapter IV, I relieve this limitation by modelling binary traits *via* unobserved normal variables, i.e. latent liabilities. This approach could be used to extend driftsel for many kinds of non-normal traits, as has been done for other quantitative-genetic software packages (e.g. Hadfield & Nakagawa 2010).

Thirdly, while I find the  $S$  and  $H$  statistics as very natural starting points to detect natural selection, other possibilities would surely also exist. The  $S$  and

$H$  statistics are based on the posterior distribution of mean additive genotypes  $\mathbf{a}^p$ , yielding intuitively from the question ‘are these populations different, on average’. Another line of reasoning would be to compare the patterns of variation among local populations, i.e. the local  $\mathbf{G}$  matrices (Calsbeek & Goodnight 2009; Chapuis *et al.* 2008; Martin *et al.* 2008) which are also known to change in response to natural selection (Calsbeek & Goodnight 2009; Steppan *et al.* 2002). It should also be noted here, that the numerical values of  $S$  and  $H$  have a similar interpretation as classical test statistics (though not identical, given that our analyses are based on Bayesian statistics), with certain logical limitations: For example,  $S = 0.96$  is interpreted as a sign of natural selection, acknowledging a false positive rate of 0.04. However, this does not imply that natural selection has taken place with a posterior probability of 0.96. The posterior probability of natural selection cannot be calculated, because there is no explicit model for natural selection in this framework.

Mutation has been ignored when deriving Eq. 15. This can be a major problem in some study systems. For example, let us consider populations with a long history of reproductive isolation. On the limit of a very high random drift,  $\boldsymbol{\theta}^p$  tends to a unit-diagonal matrix as discussed in Chapter I, and consequently

$$\mathbf{a}^p \rightarrow \text{MVN}(\boldsymbol{\mu} \otimes \mathbf{I}, 2\mathbf{G}^A \otimes \mathbf{Id}) \quad (\text{Eq. 16})$$

where  $\mathbf{Id}$  is an identity matrix, so that

$$\text{Var}(a_i^p) \rightarrow 2G_{ii}^A, \quad (\text{Eq. 17})$$

i.e. the maximal variance of population means will be twice the ancestral additive genetic variance. This maximum corresponds to a situation where all populations are clonal, and all coding loci are homozygous. Consequently, Eqs. 16 and 17 imply that there is a probabilistic limit to inter-population diversity that can derive from the ancestral population. In reality, if random genetic drift is let to operate long enough, this limit will be overcome by the accumulation new mutations. Thus, Eq. 15 may give an overly conservative picture of patterns of diversity. This also concerns our statistical tests which may yield false positives when applied to e.g. inter-species comparisons. On the other hand, the  $S$  test is known to be overly conservative in the original framework of negligible mutation rate and a relatively short demographic history (see Chapter II).

Ignoring mutation shows also in the theoretical considerations of Chapter I. Here the parameters of interest are coancestry coefficients, *'formally defined as the probability that a pair of homologous genes derive from the same allelic copy in the ancestral population'*. The admixture model and the likelihood of data are derived without mutation, and it is fairly straight forward to show that this type of model underestimates the true coancestry coefficients when applied on data with a high mutation rate (Slatkin 1991, 1995). The concept of ancestral population in itself is also problematic. This concept is needed to explain polymorphism of genes in the data as a subsample of ancestral variation. Yet, we do not ask when or where the ancestral population was, and more importantly, what was the origin of ancestral polymorphism. This is a slightly artificial frame of reference, and it is not needed in models which focus on probability of IBD instead of coancestry and explicitly model polymorphism as a result of mutation (Rousset 2004). In analogy with the concept of ancestral population, the local populations in this thesis are discrete entities, which may be a simplification in some cases. Much of present population-genetic theory concerns individuals in continuous space, and elaborate statistical methods have been developed to assign individuals to different subpopulations, or to assess the contributions of different ancestral lineages within each individual (e.g. Corander *et al.* 2008; Falush *et al.* 2003).

## 5. ACKNOWLEDGEMENTS

I wish to thank my supervisors Otso Ovaskainen and Juha Merilä for challenges and encouragement. I know that I have not always been the easiest PhD student, but yet we have been able to push the work through. At times, it has even been fun! I also wish to thank numerous field and lab biologists from Juha's group, whether co-authors or not, for having accumulated

the data and biological background used in my thesis. I have stood on the shoulders of giants – well, at least large sticklebacks.

Regarding the end product of this process, this PhD thesis, thanks are due for a few people: For Jarrod Hadfield, Mikko Sillanpää and Jukka Corander for agreeing to review this thesis, in spite of their busy schedules; and for Sami Ojanen for preparing the layout of this book.

I also wish to thank my group, the MRG, for international atmosphere and lively discussion, both scientific and informal. In general, a number of people in Viikki would deserve specific thanks and acknowledgements. I will limit this paragraph to the most obvious ones: I thank my scientific companion and room neighbour Tanjona Ramiadantsoa for his constant good moods and sarcastic humour. *"At some point, I think that this girl was watching at us, but I didn't want to give the wrong impression, so I would eventually go home."* I also wish to thank Veera Norros for her good parties and delicious cooking. If some occasion needs to be highlighted from the course of my PhD studies, that would most probably be the fluctuating populations' workshop with Tanjona, Henna and Ayco in Norway in 2012. I can still feel the dehydration from the three-hour sauna sessions!

Likewise, a number of personal friends definitely deserve thanks, but I will cut the list short and thank Tatu Westling and Rasmus Ahvenniemi: You have fed me, when I have been hungry; you have lent me, when I have been broke; you have even shared my career angst with me. That's what true friends do.

Finally, I wish to thank my parents for encouragement and support.

Markku

<b>Glossary: some concepts in population genetics.</b>	
ancestral population	a panmictic reference population at some time in distant past; all local populations are thought to have split from the ancestral population
coalescent	the pedigree for a set of gene copies seen as a random sampling backward in time, from generation to another
coalescence time	expected no. of generations since the most recent common ancestor (MRCA) for a pair of homologous genes; e.g. $t_B$ 'between populations' and $t_W$ 'within populations'
coancestry coefficient	the probability that a pair of homologous genes derive from the same allelic copy in an ancestral population
probability of identity by descent (IBD)	the probability that two homologous genes have not mutated into different genotypes since their most recent common ancestor
identity by state (IBS)	the event that two homologous genes are of the same genotype; IBD implies IBS
limit of low mutation rate	the case where mutation rate is negligible compared to the effects of other evolutionary forces; often investigated mathematically by taking the limit $\mu \rightarrow 0$ , e.g. $\lim_{\mu \rightarrow 0} F_{ST} = \frac{t_B - t_W}{t_B}$
<b>D</b>	the variance-covariance matrix of mean additive genotypes (breeding values) among local populations
$F_{ST}$	Wright's fixation index; an index in $[0, 1]$ measuring genetic differentiation of subpopulations. 0 implies panmixia, and 1 implies clonal populations, each homozygous for a different genotype
<b>G</b>	the variance-covariance matrix of additive genetic effects (breeding values); can refer to a local population or the whole metapopulation, depending on the context
$Q_{ST}$	index of quantitative-genetic differentiation; interpretation and scale similar to $F_{ST}$

## REFERENCES

- Arnold SJ, Bürger R, Hohenlohe PA, Ajie BC, Jones AG (2008) Understanding the evolution and stability of the G-matrix. *Evolution* **62**, 2451-2461.
- Brommer JE (2011) Whither P(st)? The approximation of Q(st) by P(st) in evolutionary and conservation biology. *Journal of Evolutionary Biology* **24**, 1160-1168.
- Calsbeek B, Goodnight CJ (2009) Empirical comparison of G matrix test statistics: finding biologically relevant change. *Evolution* **63**, 2627-2635.
- Chapuis E, Martin G, Goudet J (2008) Effects of selection and drift on G matrix evolution in a heterogeneous environment: a multivariate Q(st)-F(st) test with the freshwater snail *Galba truncatula*. *Genetics* **180**, 2151-2161.
- Chenoweth SE, Blows MW (2008) Q(st) meets the G matrix: the dimensionality of adaptive divergence in multiple correlated quantitative traits. *Evolution* **62**, 1437-1449.

- Corander J, Marttinen P, Siren J, Tang J (2008) Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. *BMC Bioinformatics* **9**, 539.
- Darwin C (1859) *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, London.
- Dias PC, Blondel J (1996) Local specialization and maladaptation in the Mediterranean blue tit (*Parus caeruleus*). *Oecologia* **107**, 79-86.
- Edelaar P, Burraco P, Gomez-Mestre I (2011) Comparison between Q(st) and F(st) - how wrong have we been? *Molecular Ecology* **20**, 4830-4839.
- Egea R, Casillas S, Barbadilla A (2008) Standard and generalized McDonald-Kreitman test: a website to detect selection by comparing different classes of DNA sites. *Nucleic Acids Research* **36**, W157-162.
- Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567-1587.
- Feller W (1950) *An Introduction to Probability Theory and Its Applications*, 3rd edn. John Wiley & Sons, Inc., Hoboken, NJ.
- Flint J, MacKay TFC (2009) Genetic architecture of quantitative traits in mice, flies, and humans. *Genome Research* **19**, 723-733.
- Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* **180**, 977-993.
- Gelman A, Carlin JB, Stern HS, Rubin DB (2004) *Bayesian Data Analysis*, 2nd edn. Chapman and Hall/CRS, Boca Raton, FL.
- Grant PR, Grant BR (1995) Predicting microevolutionary responses to directional selection on heritable variation. *Evolution* **49**, 241-251.
- Hadfield JD, Nakagawa S (2010) General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of Evolutionary Biology* **23**, 494-508.
- Hanski I, Kuitunen J (1986) Shrews on small islands: epigenetic variation elucidates population stability. *Holarctic Ecology* **9**, 193-204.
- Hartl DL, Clark AG (2007) *Principles of Population Genetics*, 4th edn. Sinauer Associates, Sunderland, MA.
- Hastings WK (1970) Monte-Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97-8.
- Herczeg G, Gonda A, Merilä J (2009) Evolution of gigantism in nine-spined sticklebacks. *Evolution* **63**, 3190-3200.
- Hill WG (2010) Understanding and using quantitative genetic variation. *Philosophical Transactions of the Royal Society B-Biological Sciences* **365**, 73-85.
- Hill WG, Goddard ME, Visscher PM (2008) Data and theory point to mainly additive genetic variance for complex traits. *Plos Genetics* **4**, e1000008.
- Kimura M (1968) Evolutionary rate at the molecular level. *Nature* **217**, 624-626.
- Kimura M (1983) *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge.
- Kingsolver JG, Hoekstra HE, Hoekstra JM, et al. (2001) The strength of phenotypic selection in natural populations. *American Naturalist* **157**, 245-261.
- Kolmogorov A (1933) *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Julius Springer, Berlin.
- Kronholm I, Loudet O, de Meaux J (2010) Influence of mutation rate on estimators of genetic differentiation - lessons from *Arabidopsis thaliana*. *BMC Genetics* **11**, 1471.
- Lande R (1979) Quantitative-genetic analysis of multivariate evolution, applied to brain - body size allometry. *Evolution* **33**, 402-416.
- Lande R, Arnold SJ (1983) The measurement of selection on correlated characters. *Evolution* **37**, 1210-1226.
- Le Corré V, Kremer A (2012) The genetic differentiation at quantitative trait loci under local adaptation. *Molecular Ecology* **21**, 1548-1566.
- Leinonen T, McCairns RJ, O'Hara RB, Merilä J (2013) Q(st)-F(st) comparisons: evolutionary and ecological insights from genomic heterogeneity. *Nature Reviews Genetics* **14**, 179-190.
- Leinonen T, O'Hara RB, Cano JM, Merilä J (2008) Comparative studies of quantitative trait and neutral marker divergence: a meta-analysis. *Journal of Evolutionary Biology* **21**, 1-17.
- Lynch M, Walsh B (1998) *Genetics and analysis of quantitative traits*. Sinauer Associates Incorporated, New York, NY.
- MacKay TFC (2004) The genetic architecture of quantitative traits: lessons from *Drosophila*. *Current Opinion in Genetics & Development* **14**, 253-257.
- Martin G, Chapuis E, Goudet J (2008) Multivariate Q(st)-F(st) comparisons: A neutrality test for the evolution of the G matrix in structured populations. *Genetics* **180**, 2135-2149.
- Merilä J (2013) Nine-spined stickleback (*Pungitius pungitius*): an emerging model for evolutionary biology research. *Annals of the New York Academy of Sciences* **1289**, 18-35.
- Merilä J, Crnokrak P (2001) Comparison of genetic differentiation at marker loci and quantitative traits. *Journal of Evolutionary Biology* **14**, 892-903.
- Miller JR, Wood BP, Hamilton MB (2008) F(st) and Q(st) under neutrality. *Genetics* **180**, 1023-1037.



- Muenchen B (2013) The Popularity of Data Analysis Software. <http://r4stats.com/articles/popularity/>.
- Nicholson G, Smith AV, Jonsson F, et al. (2002) Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **64**, 695-715.
- O'Hara RB, Merilä J (2005) Bias and precision in Q(st) estimates: Problems and some solutions. *Genetics* **171**, 1331-1339.
- Ohta T (1973) Slightly deleterious mutant substitutions in evolution. *Nature* **246**, 96-98.
- Ohta T (1992) The nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics* **23**, 263-286.
- Ovaskainen O, Cano JM, Merilä J (2008) A Bayesian framework for comparative quantitative genetics. *Proceedings of the Royal Society B-Biological Sciences* **275**, 669-678.
- Pulliam HR, Danielson BJ (1991) Sources, sinks, and habitat selection - a landscape perspective on population-dynamics. *American Naturalist* **137**, S50-S66.
- R Development Core Team (2012) R: A language and environment for statistical computing. R foundation for statistical computation, Vienna.
- Ridley M (2004) *Evolution*, 3rd edn. Blackwell Science, Malden, MA.
- Roff DE (1997) *Evolutionary Quantitative Genetics*, 1st edn. Chapman & Hall, New York, NY.
- Rousset F (2002) Inbreeding and relatedness coefficients: what do they measure? *Heredity* **88**, 371-380.
- Rousset F (2004) *Genetic Structure and Selection in Subdivided Populations* Princeton University Press, Princeton, NJ.
- Schlötterer C (2000) Evolutionary dynamics of microsatellite DNA. *Chromosoma* **109**, 365-371.
- Selkoe KA, Toonen RJ (2006) Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology Letters* **9**, 615-629.
- Shikano T, Shimada Y, Herczeg G, Merilä J (2010) History vs. habitat type: explaining the genetic structure of European nine-spined stickleback (*Pungitius pungitius*) populations. *Molecular Ecology* **19**, 1147-1161.
- Slatkin M (1991) Inbreeding coefficients and coalescence times. *Genetical Research* **58**, 167-175.
- Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**, 457-462.
- Spitze K (1993) Population structure in *Daphnia obtusa*: quantitative genetic and allozymic variation. *Genetics* **135**, 367-374.
- Stephan SJ, Phillips PC, Houle D (2002) Comparative quantitative genetics: evolution of the G matrix. *Trends in Ecology & Evolution* **17**, 320-327.
- Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology* **14**, 671-688.
- Tavaré S (1984) Line-of-descent and genealogical processes, and their applications in population-genetics models. *Theoretical Population Biology* **26**, 119-164.
- Templeton AR (2006) *Population Genetics and Microevolutionary Theory*. John Wiley & Sons, Hoboken, NJ.
- Torres-Dowdall J, Handelsman CA, Reznick DN, Ghalambor CK (2012) Local adaptation and the evolution of phenotypic plasticity in Trinidadian guppies (*Poecilia reticulata*). *Evolution* **66**, 3432-3443.
- Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population-structure. *Evolution* **38**, 1358-1370.
- Weir BS, Hill WG (2002) Estimating F-statistics. *Annual Review of Genetics* **36**, 721-750.
- Whitlock MC (1999) Neutral additive genetic variance in a metapopulation. *Genetical Research* **74**, 215-221.
- Whitlock MC (2008) Evolutionary inference from Q(st). *Molecular Ecology* **17**, 1885-1896.
- Whitlock MC (2011) G'(st) and D not replace F(st). *Molecular Ecology* **20**, 1083-1091.
- Whitlock MC, Gilbert KJ (2012) Q(st) in a hierarchically structured population. *Molecular Ecology Resources* **12**, 481-483.
- Whitlock MC, Guillaume F (2009) Testing for spatially divergent selection: comparing Q(st) to F(st). *Genetics* **183**, 1055-1063.
- Wright S (1951) The Genetical Structure of Populations. *Annals of Eugenics* **15**, 323-354.
- Wright S (1969) *Evolution and the Genetics of Populations II: The Theory of Gene Frequencies*. Chicago University Press, Chicago, IL.